Exploitation of Clustering Techniques in Transactional Healthcare Data

NAEEM AHMED MAHOTO*, FAISAL KARIM SHAIKH**, AND ABDUL QADIR ANSARI***

RECEIVED ON 28.09.2013 ACCEPTED ON 09.01.2014

ABSTRACT

Healthcare service centres equipped with electronic health systems have improved their resources as well as treatment processes. The dynamic nature of healthcare data of each individual makes it complex and difficult for physicians to manually mediate them; therefore, automatic techniques are essential to manage the quality and standardization of treatment procedures. Exploratory data analysis, pattern-analysis and grouping of data is managed using clustering techniques, which work as an unsupervised classification. A number of healthcare applications are developed that use several data mining techniques for classification, clustering and extracting useful information from healthcare data. The challenging issue in this domain is to select adequate data mining algorithm for optimal results. This paper exploits three different clustering algorithms: DBSCAN (Density-Based Clustering), agglomerative hierarchical and k-means in real transactional healthcare data of diabetic patients (taken as case study) to analyse their performance in large and dispersed healthcare data. The best solution of cluster sets among the exploited algorithms is evaluated using clustering quality indexes and is selected to identify the possible subgroups of patients having similar treatment patterns.

Key Words: Clustering Techniques, Data Mining, Healthcare Applications, Diabetes.

1. INTRODUCTION

Healthcare management profits the solutions offered by IT (Information Technology) for the betterment of healthcare services as well as reducing unnecessary costs and time. The key issue in healthcare management is to properly cure the deadly diseases at appropriate time and to adopt proper procedures. Clinical pathway (treatment course undertaken by a patient for a certian disease) plays vital role in the health services, especially, to manage the quality and the standardization of treatment procedures [1-3]. Since, individual patient has its own family history and state of the symptoms, therefore, diversity in the clinical pathways is often observed. This dynamic nature of clinical pathways is essential for the physicians to carefully decide the diagnostic procedures. In addition diversity in clinical pathways also makes the treatment procedure as an evolving process. However, dealing with long and complex clinical pathways require technological solutions where manually interventions are very much costly in terms of time consumption and error-free analysis as well as difficult to handle properly and efficiently in less time.

* Assistant Professor, Department of Sotware Engineering, Mehran University of Engineering & Technology, Jamshoro.
** Associate Professor, Department of Telecommunication Engineering, Mehran University of Engineering & Technology, Jamshoro.
*** Senior Manager, Multimedia and Broadband, Pakistan Telecommunication Company Limited, Hyderabad.

The adoption of data mining techniques in healthcare data has greatly taken attention for last many years [4-8]. A number of healthcare applications are proposed to improve healthcare services and management tasks in literature, for example, the work in [6] analysed colon cancer data to detect medical pathways followed by patients and advised to enrich the existing care guidelines. Similarly, application of closed frequent itemset techniques in constructing clinical pathways in a reverse engineering approach is proposed in [4]. Furthermore, the use of high-tech instruments in healthcare services have provided an immense data, which can be beneficial for the management as well as care procedures. However, investigating all records present in the healthcare data is not straight and easy task, therefore, sophisticated and efficient techniques are needed to get useful information from such a large data. Among the most popular and robust data mining techniques, association rule mining [9], sequential pattern mining [10], clustering [11] and classification [11] are well-known and frequently exploited in several knowledge discovery processes including healthcare data [4,6,12-13]. The problem in healthcare data is its high dimensionality, complex and disperse nature, which makes it even more complicated, thus it is an open issue and challenging task to uncover useful information from healthcare data.

Clustering is considered as unsupervised classification technique, which aims at identifying the similar groups of entities available in given dataset. The entities in a subgroup (i.e. cluster) formed in clustering process are homogeneous to members of its own subgroup and heterogeneous to members of other subgroups (i.e. clusters) [11]. Although clustering has ability to distinguish identical entities in a given dataset, but while dealing healthcare data, the process becomes more complex and difficult to obtain clusters, due to diversified and sparse nature of healthcare data. This study proposes an approach to cluster patients of similar clinical pathways using different clustering technique on a real healthcare dataset (data is provided by healthcare aganecy). The reason behind the clustering techniques applied on healthcare data is that different groups of entities in certian medical treatment is highly complicated issue because each individual has its own medical hisotry. Therefore, clustering algorithms working on different distance methods are exploited. For example, DBSCAN algorithm [14] works on density based methods, K-mean [15] on partitioning method and agglomerative hierarchical clustering [16] on hierarchical methods. The experiments are performed on real healthcare data for all considered clustering algorithms. The transactional healthcare data is firstly transformed into vectors representing physical diagnostic examinations of each patients, then different clustering techniques are applied to achieve subgroups present in the data. The clustering results are finally evaluated by means of quality indexes.

The rest of the paper is organized as follows. Section 2 presents literature review of data mining techniques used in healthcare data. The approach used to cluster healthcare data is described in Section 3, the results obtained are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. DATA MINING IN HEALTHCARE DATA: A LITERATURE REVIEW

Data mining is widely used in several domains for identifying, detecting and analysing data that is beneficial for future predictions, improving current processes and managing resources in optimal way. In [17] benefits and limitations of IT in healthcare, and recommendations, as need of the hour to achieve desired outcomes, are reported. An integrated environment, which has provision of personalized healthcare services meeting the user specifications is described in [7]. The safety of patients is one of the essential factors in HIT (Healthcare Information Technology) systems. Furthermore, HIT systems that lack in proper design and implementation may lead towards severe errors. Hence, enhanced approaches are recommended for HIT to deliver safer HIT services [8]. The work in [18] reported that the use of HIT system significantly improves service quality as well as reduces cost.

Three Algorithms: ANN (Artificial Neural Networks), decision trees (machine learning methods) and logistic regression (statistical method) are applied in prediction of breast cancer from a large dataset in [13]. The comparison between three algorithm is based on 10-fold cross-validation methods, and concluded decision tree (C5) performed the best in accuracy, followed by ANNs (multi layered perceptron architecture) as second best and finally the least accuracy was observed in statistical method logistic regression.

An approach for association rules in medical databases (called fuzzy association rules), that improves semantics of rules in both antecedent and consequent by means of fuzzy sets is reported in [19] and thus useful information with better semantics are obtained. A model has been developed in [20] to process, manage and analyse the information entities and their relationships in large-scale clinical data. Moreover, several OLAP (Online Analytical Processing) and data mining tasks are done to discover knowledge from TCM (Traditional Chinese Medicine) data. An expert system, developed in [21], generates implication rules based on knowledge base to reduce the time focusing on breast cancer trials. The study about exploring risk factors for perinatal morbidity and mortality is discussed in [22] exploiting neural network and decision tree (C5), and reported drinking and smoking along with other attributes related to mother being the important risk factors. The work in [23] recommended involvement of patients and experts in the design of healthcare applications.

Clustering is greatly exploited in exploratory data analysis, pattern-analysis and grouping. A number of approaches are reported in literature for clustering healthcare data with respect to different aspects. For example, K-means [15] algorithm is applied in [24] to cluster healthcare data, after transformation of binary data into real data. The transformation of data is made possible using Linear Wiener Transformation, which is normally used in noise filtering and is statistical transformation. The k-means cluster analysis has been incorporated into a methodology MCA (Multiple Correspondence Analysis), which investigates the characteristics of the people who use multiple healthcare resources in [25]. The proposed methodology in [25] helps in finding attribute clustering in an optimal way. Furthermore, v-fold cross-validation is exploited in k-means cluster analysis to analyse the socioeconomic and demographic characteristics of the people in the considered dataset, which are related to health care choices. The categorization of diabetic patients in [26] is carried out by hybrid model - three steps in cascaded fashion. Firstly, incorrect classified instances are identified and removed by k-means clustering approach followed by second step, where GA (Genetic Algorithm) and CFS (Correlation based Feature Selection) are applied for extracting relevant features. Finally, KNN (K-Nearest Neighbor) classifier is used for classification. The hybridized k-means clustering algorithm is proposed in [12], which uses PCA (Principal Component Analysis) method on data and then k-means clustering is applied on resultant reduced data for analysing high dimentional data. The experiments have been carried out on three different datasets of UCI machine learning repository: (i) Pima Indian Diabetes dataset, (ii) Breast Cancer dataset and (iii) SPECTF Heart dataset [12]. A fuzzy clustering technique that is based on symmetry has been developed in [27] to solve microarray data. A framework of subspace clustering has been proposed in [28] to examine the clustering of patient records for chronic diseases like diabetes and stroke.

EM (Expected Maximization) is applied in [29] to detect normal and abnormal ECG (Electrocardiography) clusters form compressed ECG data. An algorithm has been proposed in [29] for the identification of cardiac abnormalities. An unsupervised clustering to cluster individual time series for categorizing the biosurveillance data is reported in [30]; though the results are potential but the procedure takes much time and produces large number of clusters. The proposed algorithm in [30]uses Markov chain Monte-Carlo simulation in a Bayesian model.

A bisecting k-prototypes algorithm that analyses and identifies risks in healthcare data in an unsupervised clustering fashion is proposed in [31]. TCRS (Tight Clustering for Rare Senses) a clustering based method is developed in [32] for detecting senses from abbreviations often used in clinical documents. The results were compared with EM algorithm, and TCRS outperformed EM algorithm on average. The work in [33] demonstrated K-means clustering data mining technique to support building analytical models of patient flow in hospitals. The work in [33] emphasized the data preparation step as important issue that impact the Quality of solutions using the data, furthermore, size of dataset matters the solutions. The authors in [34] discussed for discovering and integrating frequent sets of features from distributed databases by means of unsupervised learning (i.e. agglomerative hierarchical clustering). The distributed datasets are mined for frequent sets and then merged in to a single frequent itemset; further, hierarchical clustering is performed and finally the cluster indexing is measured for the quality of results.

This study proposes an approach for clustering transactional healthcare data and exploits a number of clustering techniques such as DBSCAN [14], k-means [15] and agglomerative hierarchical [16] to identify the subgroups of similar clinical pathways of real healthcare data of diabetic patients. The aim is to uncover the different clinical pathways of diabetic patients, which will help in further treatment procedures, such as expected severeness of certain patients for a given condition.

Further, the clustering results are compared using clustering quality indexes to analyse the behaviour of each adopted clustering technique.

3. CLUSTERING HEALTHCARE DATA

The proposed approach illustrated in Fig. 1 converts the transactional healthcare data into subgroups of similar patients having similar clinical pathways. It comprises of four main blocks. The role and importance of each block is described in the following.

3.1 Data Collection

A real transactional healthcare dataset is provided by an agency (name is kept secret due to privacy reasons). In particular, it comprises records of 6380 diabetic patients.

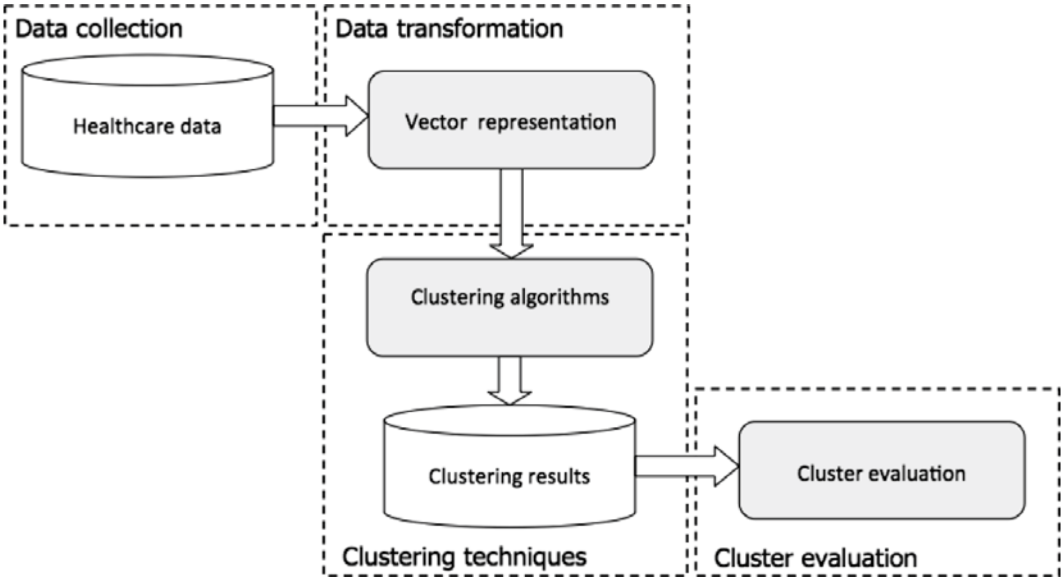

FIG. 1. PROPOSED APPROACH OF CLUSTERING HEALTHCARE DATA

The dataset is processed to transform into a format suitable for the clus- tering techniques by removing unnecessary and noisy data. For example, irrelevant data attributes like medical unit and patient's address are discarded.

3.2 Data Transformation

The data transformation phase transformed cleaned dataset into vector space model representation suitable for clustering techniques, exploited in documents clustering and text [35, 36]. Each patient, P, is considered as a vector in the term- space. The Vector, P, is represented as given in the following equation:

$$P_{ef} = (ef_1; ef_2; ef_3;..., ef_n) \qquad (1)$$

where ef_i is the frequency of i^{th} physical diagnostic examination for the given patient.

For example, consider a dataset, *D*, with two vectors or patients: $Patient_A$ and $Patient_B$ presented in Table 1. The patients, who are never diagnosed any exam (i.e. diagnostic test or physical diagnostic examination) are mentioned with zero in the vector and presence of exam is reported with its frequency for the given patient.

The rows in Table 1 represent the vectors or patients and columns are the distinct exams of the given dataset. The value “0" indicates the absence of the exam and non-zero values reports the number of occurrence of that exam for a given patient, e.g. $Patient_A$ has done thrice e_1, twice e_2, once e_4, and the same patient never did e_3 and e_5.

3.3 Clustering Techniques

Several clustering algorithms are proposed for identifying similar groups of data objects [11]. Clustering algorithms can be classified according to (i) the theory and fundamental concepts on which clustering analysis techniques are based, (ii) the ability to deal with noisy data, i.e. the sensitivity to outliers, missing or erroneous data, (iii) the a-priori knowledge on data required to correctly set the input parameters, (iv) the ability to deal with clusters of different shapes and density. These algorithms may be categorized as: (i) density-based methods, (ii) partitioning methods, (iii) model-based methods, and (iv) hierarchical-based methods.

TABLE 1. THE VECTOR SPACE MODEL REPRESENTATION OF DATASET D

Vectors	e_1	e_2	e_3	e_4	e_5
$Patient_A$	3	2	0	1	0
$Patient_B$	0	1	1	5	2

Density-based methods are less sensitive to the presence of outliers and identify non-spherical shaped clusters. These methods identify portions of the data space characterized by a high density of objects. Density is defined as the number of objects which are in a particular area of the n-dimensional space. The general strategy is to explore the data space by growing existing clusters as long as the number of objects in their neighborhood is above a given threshold. DBSCAN [14] algorithm is an example of this category.

Partitioning methods attempt to decompose a dataset of n objects into K disjoint partitions, where $k<n$. The general criterion to perform partitioning assigns objects to the same cluster when they are close, and to different clusters when they are far apart with respect to a certain metric (i.e. euclidean distance). Partitioning methods are able to find only spherical-shaped clusters, unless the clusters are well separated, and are sensitive to the presence of outliers. K-Means [15] is a popular method which belongs to this category.

Model-based methods hypothesize a mathematical model for each cluster, and then analyse the data set to determine the best fit between the model and the data. These algorithms are able to correctly take into account outliers and noise by making use of standard statistical techniques. EM [37] algorithm is one example of this category.

Hierarchical-based methods generate a hierarchical collection of clusters by means of either an agglomerative or divisive approach. The first one, which is the most common, starts with all points as singleton clusters and at each step, merge the closest pair of clusters. Different cluster proximity measures (e.g. single-link, complete-link, group average) [11] can be exploited to address the merge step. Since the output is a hierarchical collection of clusters, these methods are often used when the underlying application requires the creation of a taxonomy/hierarchy.

Each of the clustering techniques needs some computations to measure the similarity and dissimilarity between data points (i.e. patient's vectors in our study). The distance measurement or similarity between the vectors are computed by means of cosine similarity, becasue it produces three different results such as equal, opposite and independant between given two vectors. Cosine similarity is described in more details in the following.

Similarity Measurement: The clustering technique groups together similar objects of data (i.e. vectors of patients) by measuring similarity between them. In this study, cosine similarity is exploited, that ranges in [-1,0,+1] values. The -1 value indicates entirely opposite patients, zero value shows both patients as independent of each other and +1 value describe the exactly similar patients [11]. Mathematically cosine similarity is computed as given in the Equation (2):

where *x* and *y* represents the vectors of patients. The *x.y* is the dot product of both vectors. For example, consider two vectors shown in Table 1, which represent frequency of exams (i.e. physical diagnostic examinations) in vector format for the patients: $Patient_A$ and $Patient_B$. The frequency of each exam is the value of the exam done by the given patient (i.e. $Patient_A$ has done 3 times e_1). The cosine Similarity of both patients $Patient_A$ and $Patient_B$ is computed in Equation (3):

$$Co\sin\left(x.y\right) = \frac{x.y}{\sqrt{x.y}\sqrt{y.y}} \tag{2}$$

The cosine value 0.32998 describes that both patients are less similar (i.e. their diagnostic physical examinations are more different, since the values is non-zero and near to zero).

In this study, clustering algorithms are used belonging to different categories as mentioned earlier (Section 3.3). In other words, the proposed approach allows to use different clustering algorithms for clustering transactional healthcare data. The description of each algorithm used in the approach is reported in the following.

$$Co\sin e\left(Patient_A, Patient_B\right) = \frac{7}{\sqrt{15}\sqrt{30}} = 0.32998 \tag{3}$$

3.3.1 DBSCAN Algorithm

The clustering process is non-static in nature, since it exploits DBSCAN algorithm [14], which is unsupervised (i.e. number of clusters are unknown) clustering technique. The unsupervised clustering technique does not require in advance the number of clusters; rather it dynamically groups together the closer data points into single cluster. The DBSCAN needs two parameters (i.e. Epsilon and minPoints) to cluster data points; hereafter diabetic patients. The DBSCAN also needs no any pre-information to cluster such as number of clusters. The working principle of DBSCAN is to group together patients into a cluster having similar density regions. The maximum allowed distance between two patients is defined by epsilon value where as minPoints defines minimum patients to form a cluster. The patients (i.e. data points) non-compliant the given parameters are referred to as outliers. The outliers are treated as noisy data points.

3.3.2 K-Means Algorithm

K-Means is a popular, widely used, partitioning algorithm [15]. It requires k as input parameter, representing the number of partitions (i.e. clusters) in which the dataset should be divided. k-means algorithm represents each cluster with the mean value of the objects it aggregates, called centroid and is based on an iterative procedure, preceded by a set-up phase, where k objects of the dataset are randomly chosen as the initial centroids. Each iteration performs two steps. In the first step, each object is assigned to the cluster whose centroid is the nearest to that object. In the second step centroids are relocated, by computing the mean of the objects within each cluster. Iterations continue until the k centroids do not change. K-means is effective for spherical-shaped clusters. Different cluster shapes are detected only if the clusters are well separated. Similar to other partitioning methods, K-means is sensitive to outliers and requires the a-priori knowledge of the number of clusters. It does not separate outliers, instead those non-compliant data points are included into the given number of clusters.

3.3.3 Agglomerative Hierarchical Algorithm

Agglomerative hierarchical clustering produces a collection of nested clusters arranged into a hierarchical tree [16]. Agglomerative hierarchical clustering, for brevity in rest of the sections will be termed as hierarchical clustering, starts with patients (i.e. vector of the patient in the considered dataset) as individual clusters and at every step, the closer patients are merged into a single cluster [11]. The similarity between patients is referred to as cluster similarity of distance. There are several ways to cluster the individual patients such as single-linkage, complete-linkage and average-linkage also called as proximity methods [11]. The single-linkage or minimum method considers the shortest distance between any members of two clusters. Likewise, complete-linkage or maximum method considers maximum and average-linkage consider mean distance. This algorithm similar to k-means does not isolate the outliers rather assigns them into clusters.

3.4 Cluster Evaluation

The evaluation of clustering results in order to determine the validity of cluster-set is the most difficult task, particularly, when no-prior information (i.e. true partition) is known. There are several evaluation techniques that measure the quality of clustering results, which can be in general classified as (i) Internal Indexes and (ii) External Indexes.

Internal indexes allow to evaluate the cluster set, when no any prior solution is available, whilst external indexes evaluate the cluster set based on already available solutions, it verifies the cluster set against the partial/full prior solution (i.e. true partition of the dataset). F-measure, Jaccard Index and Rand Index are examples of well known external indexes. The most adopted internal indexes are homogeneity, separation and silhouette.

In this study, since prior solution of transactional healthcare data is unavailable, hence, internal indexes are exploited to evaluate the clustering results. Moreover, the aim is to identify the real partition of the large dataset in terms of small subgroups. The quality indexes are computed against each clustering algorithm by varying different parameters, for instance, Eps and minPts in DBSCAN and number of clusters in k-means and Hierarchical algorithms.

3.4.1 Homogeneity Index

Homogeneity index [38] measures the compactness between members of a cluster and is computed as Equation (4):

$$HomogeneityC_i = \frac{2}{n\left(n-1\right)} \sum_{i=1,x\neq y}^{K} s(s.y) \tag{4}$$

where C_i is a cluster for which compactness is being computed, n and K are total number of members in a cluster C_i, $s(x,y)$ is similarity function that exploits cosine similarity to measure similarity between x and y vectors of patients of cluster C_i. The higher homogeneity value indicates better compactness and better solution of placement. An average homogeneity of each cluster shows the homogeneity of entire clustering result.

3.4.2 Separation Index

Separation index [38] evaluate the average similarity between neighbouring clusters and is computed as Equation (5):

$$SeparationC_{ij} = \frac{2}{N\left(N-1\right)-Q} \sum_{i=1,j=1,i\neq j}^{K} s(x,y_j) \tag{5}$$

where C_{ij} shows the separation between C_i and C_j , N is the total number of members in both clusters C_i and C_j. Q represents the number of combinations of cluster C_i, n is total combinations pairs of members in C_i and Q is calculated as Equation (6):

$$Q = \frac{2}{n\left(n-1\right)} \tag{6}$$

Like in homogeneity, the similarity function uses cosine similarity for measurement, x_i is the patient belonging cluster C_i and y_j is the member of cluster C_j . Separation index of a single cluster is calculated with all its neighbour clusters and average of all separation indexes of each cluster is termed as separation index of entire clustering. The smaller separation index values represents better solution.

3.4.3 Silhouette Index

The silhouette index [39] evaluates the correct placement of an object (i.e. patient in this study). The silhouette values fall in between -1 and 1, where -1 represents wrong placement of the object (patient), 1 shows better placement and zero indicates that the object (patient) is at the border of cluster. Silhouette index of an object *i* of a cluster *C* can be calculated as Equation (7):

$$Silhouette_i = \frac{b\left(x\right)-a\left(x\right)}{\max\left\{a\left(x\right),b\left(x\right)\right\}}, Silhouette_i \in\left[-1,1\right] \tag{7}$$

where *a*(*x*) is the average distance of object *i* with members of cluster *C* to whom it belongs; the distance is computed by means of cosine distance. *b*(*x*) is the smallest of average distances of its neighbour clusters. The above expression computes silhouette index of single patient of a cluster *C*, the average of all computed values of each patient represents silhouette of entire cluster *C*. Thus, the mean of each cluster describe the total silhouette of clustering results.

4. RESULTS AND DISCUSSION

The obtained clustering results and their quality evaluation are presented in this section on real transactional healthcare data of diabetic patients. RapidMiner [40] tool is exploited for clustering and evaluation of clustering results and preprocessing steps are implemented in Python programming language [41]. RapidMiner can be executed on several plateforms such as Mac OSX, Linux, and Windows. Similarity, Python programs can also be run on several plateforms.

4.1 DBSCAN Algorithm

The selection of parameters in DBSCAN algorithm is challenging task, hence to cope with the issue, k-dist graph [11] is exploited to determine the appropriate parameters (i.e. *Eps* and *minPts*). The *minPts*=30 illustrated the significant distribution of patients and indicated around 3000 as the outliers. Moreover, varying different epsilon values cluster sets have been evaluated by internal quality indexes. Table 2 reports as representative of the other experimental results, when epsilon value varies in [0.25-0.4] range. These paramters are mendatory for the DBSCAN algorithm for clustering data points (Section 3.3.1).

DBSCAN algorithm has an advantage of identifying sensitivitynature towards outlies. Thus, it clusters potentially patients into appropriate number of clusters regardless of the shape of cluster. The cluster evaluation results showed ∈=0.3 and *minPts* = 30 as the best configuration of DBSCAN parameters which produced 8 number of clusters and one cluster of outliers (Table 2). Though ∈=0.25 and ∈=0.3 both produced good quality measures but ∈=0.3 is taken as the best configuration since it also uniformly distributes the larger portion of patients into clusters, unlike with ∈= 0.25, where larger portion of the patients are declared as outliers.

4.2 K-Means Algorithm

Table 3 presents the quality indexes and patients distribution when K-means algorithm is applied by varying different number of clusters. Unlike DBSCAN that needs epsilon and *minPts* parameters and identifies outliers, the k-mean algorithm requires total number of clusters for the given data points and does not identify outliers. The issue in transaction healthcare data is how to determine the correct number of clusters. Since, it is dificult to provide total number of clusers within transactional data. Therefore, several experiments have been carried out to investigate the significant number of clusters which produce good quality index values.

4.3 Agglomerative Hierarchical Algorithm

In ordered to obtain desired number of clusters using agglomerative hierarchical algorithm, flaten clustering operator [40] is applied that expands nodes in hierarchical clustering according to node's (i.e. patient vector) distances till the desired number of clusters are achieved. The total number of clusters are given in this algorithm and this does not identify outliers like DBSCAN algorithm. Moreover, complete-linkage proximity method produced better results in comparison of other proximity methods (i.e. single-link and average-link) in this study. The possible reason behind the complete-linkage method could be the dispersed nature of healthcare data, which may merge nodes (i.e. patient vectors) in better way comparing with other proximity methods.

Table 4 reports the cluster sets and their quality index values achieved by applying hierarchical clustering algorithm. When number of clusters is made equal to 12, this algorithm produced good results in terms of quality indexes (i.e. homogeneity, separation and silhouette).

4.4 Comparison

The quality indexes of each cluster set of clustering algorithms indicate the effectiveness of DBSCAN algorithm, since DBSCAN cluster-set yields higher quality

TABLE 2. DBSCAN CLUSTER SETS AND THEIR QUALITY INDEXES WHEN MINPTS=30

		C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	Outliers
∈=0.25	Patients	3123	46	68	117	31	36	47	-	2912
	Homogeneity	0.75	0.95	0.94	0.80	0.94	0.86	0.83	-	0.49
	Separation	0.022	0.049	0.0325	0.024	0.025	0.002	0.023	-	0.021
	Silhouette	0.64	1.0	1.0	0.97	1.0	0.99	0.97	-	-0.23
∈=0.30	Patients	3778	46	68	118	31	38	47	38	2216
	Homogeneity	0.74	0.95	0.94	0.81	0.94	0.85	0.83	0.84	0.43
	Separation	0.016	0.05	0.026	0.021	0.026	0.017	0.02	0.03	0.017
	Silhouette	0.54	1.0	1.0	0.97	1.0	0.97	0.97	0.89	-0.38
∈=0.35	Patients	4276	138	47	68	130	31	47	56	1587
	Homogeneity	0.70	0.77	0.84	0.95	0.86	0.94	0.82	0.63	0.37
	Separation	0.014	0.027	0.048	0.026	0.025	0.024	0.018	0.023	0.019
	Silhouette	0.48	0.77	0.98	1.0	0.92	1.0	0.88	0.87	-0.38
∈=0.40	Patients	4725	162	47	68	31	58	-	-	1289
	Homogeneity	0.75	0.72	0.84	0.93	0.94	0.64	-	-	0.32
	Separation	0.013	0.022	0.064	0.034	0.032	0.003	-	-	0.02
	Silhouette	0.38	0.77	0.98	1.0	1.0	0.89	-	-	-0.35

index values comparing to other cluster-sets produced by hierarchical and K-means algorithms. The best results from each clustering algorithms are compared in Table 5, which contains the average values of each quality index of the best results obtained by each clustering algorithm (i.e. DBSCAN, K-means and Hierarchical). It is obvious from the Table 5 that DBSCAN algorithm produced better results in terms of quality evaluation.

TABLE 3. K-MEANS CLUSTER SETS AND THEIR QUALITY INDEXES

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}
Patients	1943	1282	297	344	1449	1035	-	-	-	-	-	-
Homogeneity	0.85	0.86	0.25	0.54	0.81	0.58	-	-	-	-	-	-
Separation	0.1	0.09	0.01	0.04	0.1	0.05	-	-	-	-	-	-
Silhouette	0.44	0.32	0.26	0.19	0.26	0.05	-	-	-	-	-	-
Patients	671	423	1248	1780	301	1255	324	378	-	-	-	-
Homogeneity	0.67	0.68	0.86	0.85	0.23	0.84	0.56	0.67	-	-	-	-
Separation	0.08	0.09	0.09	0.094	0.01	0.11	0.048	0.057	-	-	-	-
Silhouette	0.03	0.01	0.26	0.51	0.23	0.29	0.24	0.27	-	-	-	-
Patients	1216	756	346	416	175	310	406	1732	863	160	-	-
Homogeneity	0.87	0.8	0.68	0.65	0.38	0.6	0.69	0.85	0.85	0.23	-	-
Separation	0.08	0.1	0.06	0.07	0.02	0.05	0.08	0.08	0.11	0.01	-	-
Silhouette	0.31	0.097	0.31	0.01	0.48	0.28	0.025	0.5	0.19	0.07	-	-
Patients	418	987	532	404	91	750	848	343	241	201	296	1269
Homogeneity	0.64	0.88	0.87	0.7	0.34	0.8	0.85	0.69	0.31	0.78	0.57	0.82
Separation	0.06	0.09	0.11	0.09	0.01	0.1	0.11	0.06	0.016	0.077	0.058	0.10
Silhouette	0.01	0.44	0.27	0.01	0.15	0.08	0.18	0.31	0.32	-0.09	0.17	0.37

TABLE 4. AGGLOMERATIVE HIERARCHICAL CLUSTER SETS AND THEIR QUALITY INDEXES

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}
Patients	746	295	1521	97	114	3607	-	-	-	-	-	-
Homogeneity	0.198	0.86	0.69	0.34	0.51	0.82	-	-	-	-	-	-
Separation	0.04	0.05	0.05	0.01	0.02	0.04	-	-	-	-	-	-
Silhouette	-0.36	0.125	0.239	0.70	0.24	0.57	-	-	-	-	-	-
Patients	450	1521	3607	295	114	129	167	97	-	-	-	-
Homogeneity	0.16	0.69	0.83	0.68	0.51	0.72	0.60	0.34	-	-	-	-
Separation	0.03	0.04	0.03	0.06	0.03	0.05	0.03	0.02	-	-	-	-
Silhouette	-0.30	0.238	0.37	0.11	0.15	-0.17	0.73	0.66	-	-	-	-
Patients	426	1521	3607	97	295	129	167	17	7	114	-	-
Homogeneity	0.17	0.69	0.83	0.34	0.68	0.72	0.60	0.26	0.53	0.51	-	-
Separation	0.03	0.03	0.02	0.015	0.05	0.04	0.03	0.001	0.003	0.025	-	-
Silhouette	-0.33	0.22	0.37	0.61	0.11	-0.17	0.73	0.94	0.35	0.15	-	-
Patients	285	295	1521	97	114	3607	167	17	7	49	92	129
Homogeneity	0.21	0.68	0.69	0.34	0.51	0.83	0.60	0.26	0.53	0.74	0.42	0.72
Separation	0.02	0.05	0.03	0.02	0.03	0.02	0.03	0.001	0.004	0.03	0.02	0.05
Silhouette	-0.29	0.11	0.22	0.56	0.08	0.37	0.71	0.94	0.35	0.24	0.54	-0.17

The best configuration of DBSCAN (i.e. ∈= 0.3 and *minPts* ∈=30) produces mean homogeneity, separation and silhouette without outliers as 0.8625, 0.026 and 0.92 respectively, where as considering outliers as cluster, then mean homogeneity, separation and silhouette are computed as 0.814, 0.025 and 0.78 respectively. Likewise, number of clusters equal to 6 in K-means algorithm have good results in terms of quality measures and number of clusters equal to 12 in hierarchical clustering produced better results. Table 5 reports the average values of quality indexes for DBSCAN algorithm, when outliers are considered as a cluster i.e. in the mean calculation outlier values are also included. The outliers detection is one of the good aspect of DBSCAN algorithm to group together similar patients and discard the patients having entirely different treatment patterns.

TABLE 5. COMPARISON OF QUALITY INDEXES OF EACH CLUSTERING ALGORITHM

	Average Homogeneity	Average Separation	Average Silhouette
DBSCAN	0.814	0.025	0.78
K-means	0.65	0.07	0.255
Hierachical	0.54	0.026	0.3

The experiments are carried out on a Pentium 4 having 2.26 GHz Intel Core 2 Duo processor with 4 Giga Bytes of RAM. The execution time of DBSCAN is reasonable since the best configuration took 9 minutes and 7 seconds, where as hierarchical algorithm is mist costly since it took more than 20 minutes and least time is consumed by k-means algorithm as shown in Fig. 2.

The analysis of each cluster obtained from DBSCAN algorithm (when best configuration is taken into

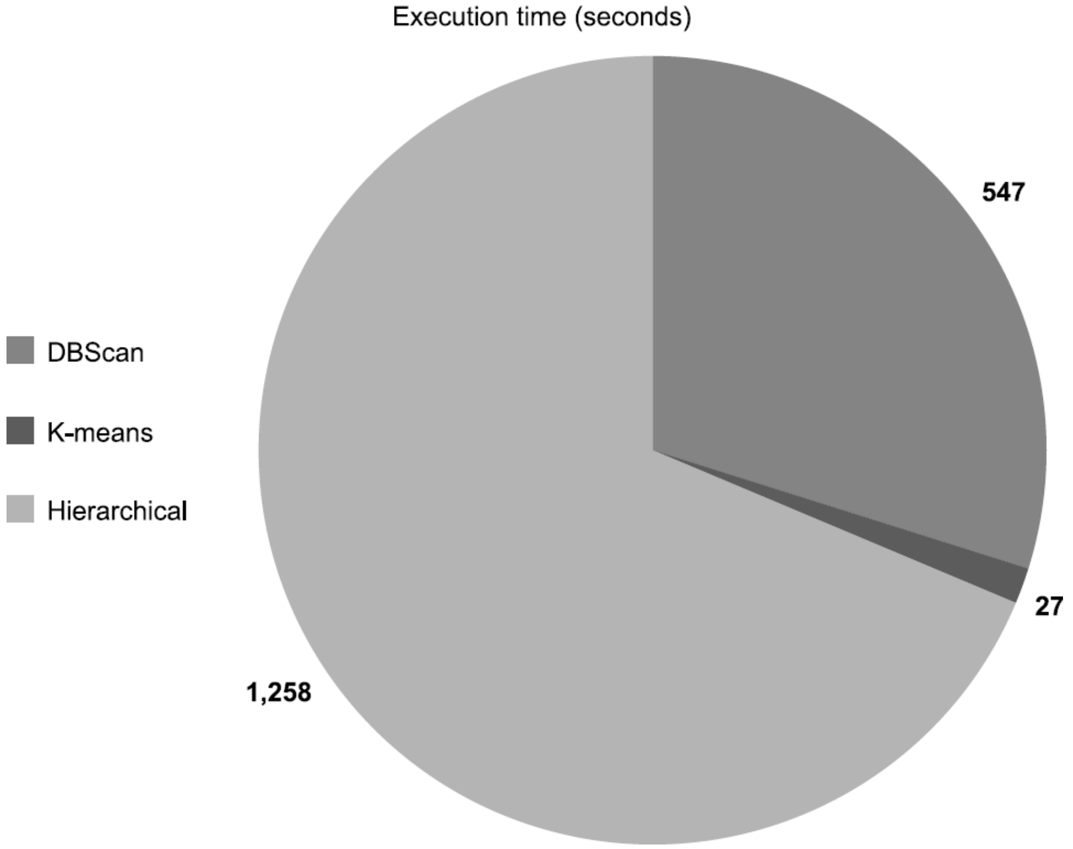

FIG. 2. PROCESSING TIME FOR EACH CLUSTERING ALGORITHM

consideration) is described in the following section. The results obtained from other clustering alogirthms are ignored, since their quality indexes were inadequate.

4.4.1 DBSCAN Algorithm Cluster Set

The best configuration of DBSCAN algorithm detected 8 cluster and an outlier. This section investigate in depth the detected cluster set to better analyse the detected subgroups of diabetic patients. Medical knowledge recommends some standard set of examinations for diabetes. For example, diagnosis tests are commonly done for diabetes include glucose level, urine test, capillary blood sample and vanous blood sample. These diagnostic exams are standard routine exams. Besides, haemoglobin, CBC (Complete Blood Count) are also performed to accurately monitor the sugar level in blood of a diabetic patient [42].

The severe complications may appear during diabetes treatment, since serious diabetic patient may also suffer from cardiovascular, liver, kidney and eye diseases [42]. Table 6 reports the standard (i.e. routine) diagnostic exams of considered transactional healthcare dataset for diabetes in detected clusters C_1;C_8 and outliers. The patients in these three clusters (i.e. C_1;C_8 and outliers) have also followed exams of cardiovascular and liver as reported in Table 7. For example, Electrocardiogram, Triglycerides and Cholesterol are cardiovascular exams, Gamma GT (Glutamyl Transpeptidase), AST (Aspartate Aminotransferase), ALT (Alanine Aminotransferase) are liver disease exams. Further Creatinine, Microalbuminuria are renal (kidney) exams and Fundus Oculi is eye exam reported in Table 8.

The other clusters: C_2;C_3;C_4;C_5;C_6 and C_7 are reported in Table 9. These clusters comprise of patients who did some specific exams. For instance, only Electrocardiogram and Fundus Oculi are performed by patients in C_2, only Fundus Oculi by C_3, only Electrocardiogram by C_5. Similarly, only Venous blood and Haemoglobin are done by C_6, both Checkup visit and Glucose level by C_7 and only one single Checkup visit is done by C_4. This behaviour shows that patients in these clusters do not follow standard routine guidelines of the diabetes. This may be because these patients are at the initial or final stages of the diabetes, since the considered dataset contains only records of one year, therefore, it may be possible these patients are at initial or final stages for the considered year or there may have been partial or erroneous data entry.

The different behaviour of the detected clusters represents the effectiveness of the approach adopted to identify the subgroups of the transactional healthcare data, which may not be visible in case of considering the dataset as a single large unit.

TABLE 6. FREQUENT ROUTINE EXAMS OF DIABETES IN DETECTED CLUSTER SET

Examinations	C_1 (%)	C_8 (%)	Outliers (%)
Checkup Visit	99	-	58
Glucose Level	100	-	71
Urine Test	100	-	45
Capillary Blood	100	-	45
Venous Blood	85	100	79
Haemoglobin	34	100	72
Complete Blood Count	2	-	38

TABLE 7. CARDIOVASCULAR AND LIVER EXAMS IN DETECTED CLUSTER SET

Examinations	C_1 (%)	C_8 (%)	Outliers (%)
Electrocardiogram	19	47	31
HDL Cholesterol	25	100	56
Cholesterol	26	100	58
Triglycerides	25	100	57
ALT	24	-	45
AST	24	-	43
Bilirubin	-	-	5
Gamma GT	4	100	35

TABLE 8. KIDNEY AND EYE EXAMS IN DETECTED CLUSTER SET

Examinations	C_1 (%)	C_8 (%)	Outliers (%)
Creatinine	10	100	38
Creatinine Clearance	15	-	19
Culture Urine	20	-	38
Microalbuminuria	11	100	40
Fundus Oculi	24	55	31

TABLE 9 .FREQUENT DIAGNOSTIC EXAMS IN DETECTED CLUSTER SET

Examinations	C_2 (%)	C_3 (%)	C_4 (%)	C_5 (%)	C_6 (%)	C_7 (%)
Checkup Visit	-	-	100	-	-	100
Glucose Level	-	-	-	-	-	100
Venous Blood	-	-	-	-	100	-
Haemoglobin	-	-	-	-	100	-
Electrocardiogram	100	-	-	100	-	-
Fundus Oculi	100	100	-	-	-	-

5. CONCLUSIONS

In this paper, cluster analysis is effectively exploited in healthcare data by applying different clustering algorithms on a real transactional healthcare data of diabetic patients. The exploited approach transformed the patients' diagnostic exams data into patient vectors, which are classified by means of clustering techniques. In particular three clustering algorithms: DBSCAN, K-means and Hierarchical have been applied. The similarity measure between the patients' vector is computed using cosine similarity.

The clustering results are evaluated using internal quality indexes for measuring the quality of detected cluster set. The results show that DBSCAN algorithm performed better than other clustering algorithms (i.e. K-means and Hierarchical) for a dense and disperse healthcare dataset. The detection of subgroups from a large transactional healthcare dataset is made possible using clustering techniques, which may not possibly be identified due to unavailability of true partitions of dataset (i.e. number of clusters). These results show the effectiveness of the approach, which may be applied in different healthcare dataset of different pathologies. Further, by detecting the potential subgroups from large healthcare dataset, healthcare management may efficiently enhance their procedures.

ACKNOWLEDGEMENTS

The authors wish to thank Giulia Bruno and Silvia Chiusano for their fruitful discussions.

REFERENCES

[1] Hunter, B., and Segrott, J.,"Re-Mapping Client Journeys and Professional Identities: A Review of the Literature on Clinical Pathways", International Journal of Nursing Studies, Volume 45, No. 4, pp. 608-625, 2008.

[2] Shen, C.P., Jigjidsuren, C., Dorjgochoo, S., Chen, C.H., Chen, W.H., Hsu, C.K., Wu, J.M., Hsueh, C.W., and Lai, M.S., and Tan, C.T.,"A Data-Mining Framework for Transnational Healthcare System", Journal of Medical Systems, pp. 1-11, 2012.

[3] Lin, F., Chou, S., Pan, S,. and Chen, Y., "Mining Time Dependency Patterns in Clinical Pathways", International Journal of Medical Informatics, Volume 62, No. 1, pp. 11-25, 2001.

[4] Baralis, E., and Bruno, G., Chiusano, S., and Domenici, V.C., Mahoto, N.A., and Petrigni, C., "Analysis of Medical Pathways by Means of Frequent Closed Sequences", Proceedings of 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems Part-III, KES'10, Berlin, Heidelberg, pp. 418-425, Springer-Verlag, 2010.

[5] Dario, A., Elena, M.B., Silvia, A.C., Mahoto, N.A., Giulia, B., and Caterina, P., “Extraction of Medical Pathways from Electronic Patient Records", Medical Applications of Intelligent Data Analysis: Research Advancements/ Rafael Magdalena-Benedito, Emilio Soria, Juan Guerrero Martı-nez, Juan Gıomez-Sanchis, Antonio Jose Serrano-Lıopez, pp. 273{289, Information Science Reference (IGI Global), 2012.

[6] Antonelli, D., Baralis, E., Bruno, G., Chiusano, S., Mahoto, N.A., and Petrigni, C., “Analysis of Diagnostic Pathways for Colon Cancer", Flexible Services and Manufacturing Journal, Volume 24, pp. 379-399, 2012.

[7] Fenza, G., Furno, D., and Loia, V., “Hybrid Approach for Context-Aware Service Discovery in Healthcare Domain", Journal of Computer and System Sciences, Volume 78, No. 4, pp. 1232-1247, 2012.

[8] Andre, W.K., David, W.B., Michael, B., Mowafa, S.H., and Elizabeth, M.B., “National Efforts to Improve Health Information System Safety in Canada, the United States of America and England", International Journal of Medical Informatics, 2013.

[9] Agrawal, R., Imielinnski, T., and Swami, A., “Mining Association Rules between Sets of Items in Large Databases", Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 207-216, New York, USA, 1993.

[10] Agrawal, R., and Srikant, R., “Mining Sequential Patterns", Proceedings of Eleventh International Conference on Data Engineering, IEEE Computer Society, pp. 3-14, Washington, DC, USA, 1995.

[11] Tan, P.N., Steinbach, M., and Kumar, V., “Introduction to Data Mining”, 2nd Edition, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2006.

[12] Rajashree, D., Debahuti, M., Amiya, K.R., and Milu, A., “A Hybridized K-means Clustering Approach for High Dimensional Dataset", International Journal of Engineering, Science and Technology, Volume 2, No. 2, pp. 59-66, 2010.

[13] Dursun, D., Glenn, W., and Amit, K., “Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods", Artificial Intelligence in Medicine, Volume 34, No. 2, pp. 113-127, 2005.

[14] Martin, E., Hans-Peter, K., Jorg, S., and Xiaowei, X., “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", pp. 226-231, AAAI Press, 1996.

[15] Juang, B.H., and Rabiner, L.R., “The Segmental K-means Algorithm for Estimating Parameters of Hidden Markov Models", IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 38, pp. 1639-1641, September, 1990.

[16] Johnson, S., “Hierarchical Clustering Schemes", Psychometrika, Volume 32, pp. 241-254, 1967.

[17] Julia, A.M., and David, W.B., “Paperless Healthcare: Progress and Challenges of an it-enabled Healthcare System", Business Horizons, Volume 53, No. 2, pp. 119-130, 2010. Special Issue on Healthcare and the Life Sciences in Transition.

[18] Indranil, R.B., and Mark, F.T., “Health Information Technology and its Impact on the Quality and Cost of Healthcare Delivery", Decision Support Systems, 2012.

[19] Miguel, D., Daniel, S., Mara. J.M.B., and Mara-Amparo, V., “Mining Association Rules with Improved Semantics in Medical Databases", Artificial Intelligence in Medicine, Volume 21, No. 13, pp. 241-245, 2001.

[20] Xuezhong, Z., Shibo, C., Baoyan, L., Runsun, Z., Yinghui, W., Ping, L., Yufeng, G., Hua, Z., Zhuye, G., and Xiufeng, Y., “Development of Traditional Chinese Medicine Clinical Data Warehouse for Medical Knowledge Discovery and Decision Support", Artificial Intelligence in Medicine, Volume 48, No. 23, pp. 139-152, 2010.

[21] Bethel, C.L., Hall, L.O., and Goldgof, D., "Mining for Implications in Medical Data", Proceedings of 18th International Conference on Pattern Recognition, Volume 1, pp. 1212-1215, IEEE Computer Society, Washington, DC, USA, 2006.

[22] Hsiang-Yang, C., Chao-Hua, C., Yao-Jung, Y., and Tung-Pi, W., "Exploring the Risk Factors of Preterm Birth Using Data Mining", Expert Systems with Applications, Volume 38, No. 5, pp. 5384-5387, 2011.

[23] Emmanuelle, B., Gillian, B., Martin, D., Charo, R., and Ifat, H.G., "Examining the Evidence of the Impact of Health Information Technology in Primary Care: An Argument for Participatory Research with Health Professionals and Patients", International Journal of Medical Informatics, Volume 81, No. 10, pp. 654-661, 2012. <ce:title>Health Information Electronic Network Systems for People Living with HIV/AIDS in Underserved Communities.

[24] Kumar, D.A., and Annie, M.C.L.C., "Clustering Dichotomous Data for Health Care", International Journal of Information, Volume 2, No. 2, 2012.

[25] Chieh-Yu, L., and Jih-Shin, L., "Mining the Optimal Clustering of Peoples Characteristics of Health Care Choices", Expert Systems with Applications, Volume 38, No. 3, pp. 1400-1404, 2011.

[26] Karegowda, A.G., Jayaram, M.A., and Manjunath, A.S., "Cascading K-means Clustering and K-nearest Neighbor Classifier for Categorization of Diabetic Patients", International Journal of Engineering and Advanced Technology, Volume 1, 2012.

[27] Sriparna, S., Asif, E., Kshitija, G., and Sanghamitra, B., "Gene Expression Data Clustering Using a Multiobjective Symmetry Based Clustering Technique", Computers in Biology and Medicine, Volume 43, No. 11, 2013.

[28] Saha, B., Pham, D.S., Phung, D., and Venkatesh, S., "Advances in Knowledge Discovery and Data Mining" Volume 7819, Lecture Notes in Computer Science, pp. 123-134, Springer Berlin Heidelberg, 2013.

[29] Fahim, S., Ibrahim, K., and Abdun, N.M., "A Clustering Based System for Instant Detection of Cardiac Abnormalities from Compressed ECG", Expert Systems with Applications, Volume 38, No. 5, pp. 4705-4713, 2011.

[30] Garrick, L.W., and William, R.H., "Unsupervised Clustering of Over-the-Counter Healthcare Products into Product Categories", Journal of Biomedical Informatics, Volume 40, No. 6, pp. 642-648, 2007.

[31] Laila, C., Jos, Z.C., and Peter, F., "Clustering-Based Methodology for Analyzing Near-Miss Reports and Identifying Risks in Healthcare Delivery", Journal of Biomedical Informatics, Volume 44, No. 5, pp. 738-748, 2011.

[32] Hua, X., Yonghui, W., Nomie, E., Peter, D.S., and Carol, F., "A New Clustering Method for Detecting Rare Senses of Abbreviations in Clinical Notes", Journal of Biomedical Informatics, Volume 45, No. 6, pp. 1075-1083, 2012.

[33] Isken, M.W., and Rajagopalan, B., "Data Mining to Support Simulation Modeling of Patient Flow in Hospitals", Journal of Medical Systems, Volume 26, No. 2, pp. 179-197, 2002.

[34] Dua, S., Dessauer, M.P., and Sethi, P., "Evaluating Cluster Preservation in Frequent Itemset Integration for Distributed Databases", Journal of Medical Systems, Volume 35, No. 5, pp. 845-853, 2011.

[35] Raghavan, V.V., and Wong, S.K.M., "A Critical Analysis of Vector Space Model for Information Retrieval", Journal of the American Society for Information Science, Volume 37, No. 5, pp. 279-287, 1986.

[36] Steinbach, M., Karypis, G., and Kumar, V., "A Comparison of Document Clustering Techniques", KDD Workshop on Text Mining, Volume 400, pp. 525-526, Boston, 2000.

[37] McLachlan, G.J., and Krishnan, T., “The EM Algorithm and Extensions, Georey, J., McLachlan, Thriyambakam Krishnan. Wiley, New York, 1997.

[38] Sharan, R., Maron-Katz, A., and Shamir, R., “Click and Expander: A System for Clustering and Visualizing Gene expression data", Bioinformatics, Volume 19, No. 14, pp. 1787-1799, 2003.

[39] Rousseeuw, P.J., “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis", Journal of Computational and Applied Mathematics, Volume 20, pp. 53-65, 1987.

[40] Rapid Miner Project, “The Rapid Miner Project for Machine Learning", Last Access on January 2013.

[41] Python Software Foundation, “Python Programming Language Official Website", Last Access on January 2013.

[42] American Diabetes Association, “Standards of Medical Care in Diabetesd-2012", Diabetes Care, Volume 35, Supplementary. 1, pp. S11S63, 2012.